自然言語処理

Natural Language Processing

第3回　2015/10/7

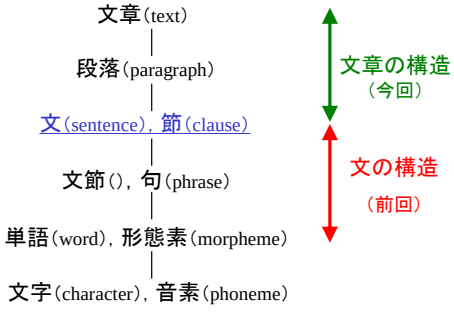
芝浦工業大学 工学部 情報工学科

杉本　徹

前回の復習，補足

- 日本語の特徴
  - 文においては述語（動詞，形容詞など）が中心
  - 述語以外の要素（主語を含む）は，述語の補足
  - 述語を修飾する言葉に対し，その言葉と述語の意味的関係を示すために，助詞（特に格助詞）を付加する
    ※参考：格文法（第7回に説明予定）
    述語とそれを修飾する要素（格要素）の関係を扱う
  - 修飾語は，被修飾語（修飾される語）より前に置く
  - 述語は，文の最後

復習：文章，文の階層構造



今日の内容

- 日本語（2）文章の構造
  - 文章の構造
  - 文章作成の技術
  - コーパス
  - 第1回課題：コーパスを使ってみよう

文章の構造

典型的な記述順序（1）

1) 序論（introduction）
   本論に入る準備をする．本論の要旨も書く
2) 本論（body）
   述べたいことを展開する
3) 結論（conclusion）
   本論のポイントをまとめて，締めくくる

典型的な記述順序（2）

1) 起　これから述べる内容を示す
2) 承　内容を発展させる
3) 転　内容を転じて別の面から述べる
4) 結　全体をまとめる

参考：漢詩の絶句

段落（Paragraph）

- 内容的に連結されたいくつかの文の集まり．全体として，ある1つの主題（topic）についてある1つのこと（考え）を述べる
- その段落が何について述べているのかを端的に述べた主題文（topic sentence）が通常存在する．主題文は段落の先頭に置かれることが多い
- 段落内の主題文以外の文は，主題文の内容を詳しく説明したり，他の段落との関係を示したりする

修辞構造（Rhetorical Structure）

- 修辞関係
  - 文間（または段落間）の意味的な関係
- 修辞構造
  - まとまりのある文章は，それぞれの文がバラバラに存在するのではなく，互いに意味的に関連し合って1つのまとまった意味的な構造を作り出している

Mann & Thompson: “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization”, *Text*, 8(3), 1988

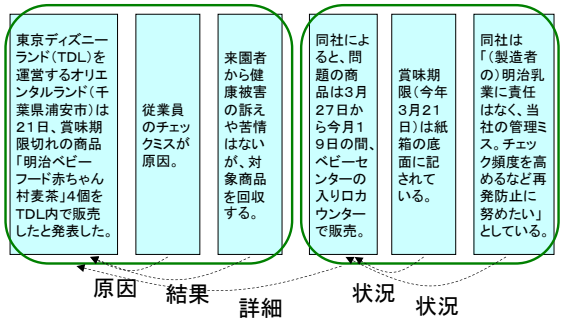
修辞構造の例

例（2008/4/21 産経新聞）：　段落

東京ディズニーランド（TDL）を運営するオリエンタルランド（千葉県浦安市）は21日、賞味期限切れの商品「明治ベビーフード赤ちゃん村麦茶」4個をTDL内で販売したと発表した。従業員のチェックミスが原因。来園者から健康被害の訴えや苦情はないが、対象商品を回収する。

同社によると、問題の商品は3月27日から今月19日の間、ベビーセンターの入り口カウンターで販売。賞味期限（今年3月21日）は紙箱の底面に記されている。同社は「（製造者の）明治乳業に責任はなく、当社の管理ミス。チェック頻度を高めるなど再発防止に努めたい」としている。

修辞構造の例



修辞関係

- 詳細（elaboration）　内容を詳細に述べる
- 原因（cause）　出来事などの原因
- 結果（result）　出来事などの結果
- 状況（circumstance）　状況の補足的な情報
- 他に，順列，並列，対比，目的，条件，根拠などの関係がある
- 修辞関係は，しばしば接続語のような手がかり語によって明示される（「したがって」，「その後」，「一方」など）．ただし明示されない場合も多い

言語使用域（Register）

- 人は状況や文脈に応じて適切な言語表現を選んで使う
- 状況の主な要素
  - 活動領域：何が起きているのか，話題は何か
  - 役割関係：話し手と聞き手の関係や社会的役割
  - 伝達様式：伝達方法，伝達媒体
- 具体例
  - 専門用語，フォーマルな言葉
  - 敬語，難しい言葉，易しい言葉
  - 書き言葉，話し言葉　など

文章作成の技術

目的

- 皆さんが技術文書を作成する際に注意すべきポイントについて考えるきっかけとなる
- 技術文書
  - 授業のレポート
  - 卒業論文，卒論概要（レジメ）
  - 仕事の報告書，計画書，企画書
  - 仕様書，取扱説明書

  など

参考文献

[1] 木下是雄：「理科系の作文技術」，中公新書，1981

[2] 本多勝一：「日本語の作文技術」，講談社，2005

[3] 大野晋：「日本語練習帳」，岩波新書，1999

[4] 酒井聡樹：「これからレポート・論文を書く若者のために」，共立出版，2007

前ページの参考文献に書かれている内容

	文献[1]	文献[2]	文献[3]	文献[4]
文書に書くべき内容（アイデアや情報の収集・整理法）	2章	ー	ー	第2部
文章の構成法（段落構成など）	3〜5章	7章	4章	第3部2章
文の正しい書き方	5〜8章	2〜6章	1〜3章	第3部3章

逆茂木型の文をなるべく避ける [1]

- 日本語は前から後に修飾するので，木構造をバラバラの葉から根に向かってたどって行くような逆茂木型の文になりやすい．複雑になると分かりにくくなるので注意する．
- 例：

コンピュータのソフトウェアには，ハードウェアを抽象化したインタフェースを提供するOSと，ユーザが実行したい機能を直接提供するアプリケーションの2種類がある．

↓

コンピュータのソフトウェアには，OSとアプリケーションの2種類がある．OSはハードウェアを抽象化したインタフェースを提供する．アプリケーションは，ユーザが実行したい機能を直接提供する．

長い修飾語は前に，短い修飾語は後に

- どの文が分かりやすいか考えてみよう

1. 太郎が駅前の店で花子が読んだ本を見つけた.
2. 太郎が花子が読んだ本を駅前の店で見つけた.
3. 駅前の店で太郎が花子が読んだ本を見つけた.
4. 駅前の店で花子が読んだ本を太郎が見つけた.
5. 花子が読んだ本を太郎が駅前の店で見つけた.
6. 花子が読んだ本を駅前の店で太郎が見つけた.

正しい文を書くためのコツ

- 正しい文を書く
  - 誤字，脱字，言葉の誤用に注意 ⇒ 念入りに見直し（推敲）
  - 文の全体的なつながりがおかしくないか確認する（重複表現の排除，呼応表現など）
  - 句読点や括弧類にも注意を払う（種類，分量）
- 読みやすい文を書く
  - 1文の長さが長くなったら，複数の短い文に分割
  - 「である」調で書く．口語表現やくだけた言葉は使わない
  - かなと漢字の分量バランスに配慮（漢字を使いすぎない）
- 明確な文を書く
  - 長文の係り受け，主語の省略，指示語などの曖昧性に注意
  - 主観的な言葉はなるべく使わない（「思う」，「とても」，「当然」，「なってしまう」など）

コーパス

コーパス（Corpus）

- 実際に書かれた言葉や話された言葉を集めた大規模な言語データ
- コーパスの種類
  - 生コーパス：収集したままの文章データ（付加情報なし）
  - タグ付きコーパス：分析に役立つ情報（単語の品詞や読み，構文・意味構造など）を生コーパスに付加したもの
  - 均衡コーパス：複数のジャンルの文章からサンプリングして統合
  - 個別用途のコーパス：学習者コーパス，模擬対話コーパスなど
  - パラレルコーパス：複数の言語間の対訳データ
  - 音声コーパス：録音した音声データ（とそれを文字化したデータ）

コーパスの例

名称	規模	内容
新聞記事データ	数千万語／年	毎日，読売，朝日など
現代日本語書き言葉均衡コーパス（BCCWJ）	約1億語	複数ジャンルの文章を収録．形態素情報付き
日本語話し言葉コーパス	約750万語	学会講演660時間分．形態素情報付き
EDRコーパス	約20万文	新聞，雑誌など．形態素・構文・意味情報付き
日英新聞記事対応付けデータ	約18万文	読売新聞とDaily Yomiuriの記事内の和文と英文の対応付け
Penn Treebank	約100万語	Wall Street Journal の記事に品詞情報・構文情報を付加
参考：Google N-gramデータ	約2550億語	Web上の文から抽出した単語n-gramデータ（後日説明）

タグ付きコーパスのデータ例（BCCWJ）

```
<sentence>
 <LUW B="B" SL="v" l_lemma="公共工事請け負い金額" l_lForm="コウキョウコウジウケオイキンガク" l_wType="混" l_pos="名詞-普通名詞-一般" >
  <SUW lemma="公共" lForm="コウキョウ" wType="漢" pos="名詞-普通名詞-一般" pron="コーキョー">
   公共
  </SUW>
  <SUW lemma="工事" lForm="コウジ" wType="漢" pos="名詞-普通名詞-サ変可能" pron="コージ">
   工事
  </SUW>
 （略）
 </LUW>
（略）
</sentence>
```

パラレルコーパスのデータ例（日英新聞記事対応データ）

<T ID="19930710E1TDY03C000010/21" NM="1-1" SCORE="1.710796165">

<J>欧州は、エディンバラにおいて合意され、コペンハーゲンにおいて強化された成長イニシアチブを精力的に実行しつつある。</J>

<E>Europe is carrying out vigorously the Growth Initiative agreed in Edinburgh and strengthened in Copenhagen. </E>

</T>

<T ID="19930710E1TDY03C000010/71" NM="1-1" SCORE="1.539716548">

<J>我々は、ロシアの経済発展にとって、改善された市場アクセスが重要であることを認識する。</J>

<E>We recognize the importance of improved market access for economic progress in Russia. </E>

</T>

コーパスの利用

- 言語学的な分析
  - 単語の出現頻度，共起頻度，使用される文脈の調査など（コーパス言語学）
- 統計的な自然言語処理
  - コーパスにタグ付けされた品詞情報や構文情報を正解データとして，それらの情報を生コーパスから自動的に求められるようにする（⇒ 機械学習やパターン認識に基づく形態素解析・構文解析）
  - 文法規則・翻訳規則・意味的関係などの言語知識の自動構築

現代日本語書き言葉均衡コーパス

（BCCWJ: Balanced Corpus of Contemporary Written Japanese）

- 現代日本語の書き言葉の全体像を把握するために国立国語研究所が開発し，2011年に公開
- 多様なジャンルの文章を含む
  - 書籍，雑誌，新聞，ベストセラー，知恵袋，法律，国会会議録，広報紙，教科書，韻文，白書，ブログ
- 形態素情報（品詞，読みなど）を付加
- Web上の検索ツール公開
  - 少納言，中納言（KWICコンコーダンサ）
  - NINJAL-LWP for BCCWJ（レキシカルプロファイラ）

NINJAL-LWP for BCCWJ（略称 NLB）の使いかた

1. Webブラウザで http://nlb.ninjal.ac.jp/ にアクセス
2. 使い方の説明を読む
3. 「検索を開始する」をクリック
4. 調べたい単語をタイプ入力し，「絞り込み」をクリック
5. 見出し，読みなどが表示されるのでクリック
6. 少し待つと，構文パターンごとの出現頻度の表が表示される
7. ここで「グループ別」の右側にある「基本」タブを選択すると，今調べている単語の文章ジャンル別の出現回数とPMW（Per Million Words, 100万語あたりの出現回数）が表示される

NLB の使いかた（続き）

8. 「グループ別」または「パターン頻度順」タブに戻り，調べてみたい構文パターンを1つ選択する
9. 画面中央の表に，コロケーションごとの出現回数が表示される
10. コロケーションを1つ選ぶと，画面右側の表に，該当する文のデータが表示される
11. この表の上部には，そのコロケーションの文章ジャンル別のPMWが示されている

- 「？」をクリックすると使い方の説明が表示される

使用例：「メール」という語の使われ方を調べる

- NLB で「メール」を検索して分かること
  - 「基本」タブの情報から，「メール」という語は知恵袋や広報紙，ブログ，雑誌でよく使われている
  - 「メールを…」という構文パターンで検索すると，多くの文章ジャンルでは「メールを送る」という言い方が最も多いが，教科書では「メールを出す」が多い
  - 「名詞＋メール」という構文パターンでは「ファーストメール」という表現の頻度が高いが，これはすべて知恵袋からのデータである

第1回課題
コーパスを使ってみよう

課題1：単語の出現頻度調査

- NLBを用いて，次の3つの文章ジャンルのうちのどれか1つでの使用頻度（PMW）が他の2つでの使用頻度（PMW）よりも高い単語の例を見つけなさい
  - ジャンル1：国会会議録
  - ジャンル2：教科書
  - ジャンル3：ブログ
- 具体的には，
  - a) 議事録でのPMWが教科書およびブログでのPMWより高い単語
  - b) 教科書でのPMWが議事録およびブログでのPMWより高い単語
  - c) ブログでのPMWが議事録および教科書でのPMWより高い単語

  の例をそれぞれ最低1個ずつ見つけて解答してください

課題2：単語の使われ方調査

- 自分で自由に単語を1つ選んで，NLBを用いて，その単語の使われ方を調査しなさい
- 調査の方法は，前述の「メール」の使われ方調査の例を参考にして，
  - いくつかの構文パターンに対して，特徴的なコロケーションを調べてまとめる
  - 例えば自分の知りたい語 X が名詞の場合，「Xを…」，「名詞＋X」，「名詞のX」などの構文パターンにおいて特徴が出やすいと思われる
  - 文章ジャンルによる違いを考察するのもよい

レポートの形式，期限，提出方法

- 形式：電子ファイル（MS Word の doc/docx，またはPDF）
- 内容構成：レポートの内容に，下記の①②③を（それぞれの課題について）含めること
  - ①調査方法の説明
  - ②調査結果（必要に応じて表やグラフを用いる）
  - ③結果についての考察
- 期限：10月20日（火）
- 提出方法：電子メールで杉本宛（sugimoto@～）

お知らせ：杉本研のプレゼミ

1） 9月30日（水）5限：研究分野・研究室紹介
2） 10月 7日（水）5限：研究分野・テーマ紹介

場所：503教室

☆ 自然言語処理の研究に興味のある人は
ぜひご参加ください！ 大学院進学希望者歓迎！